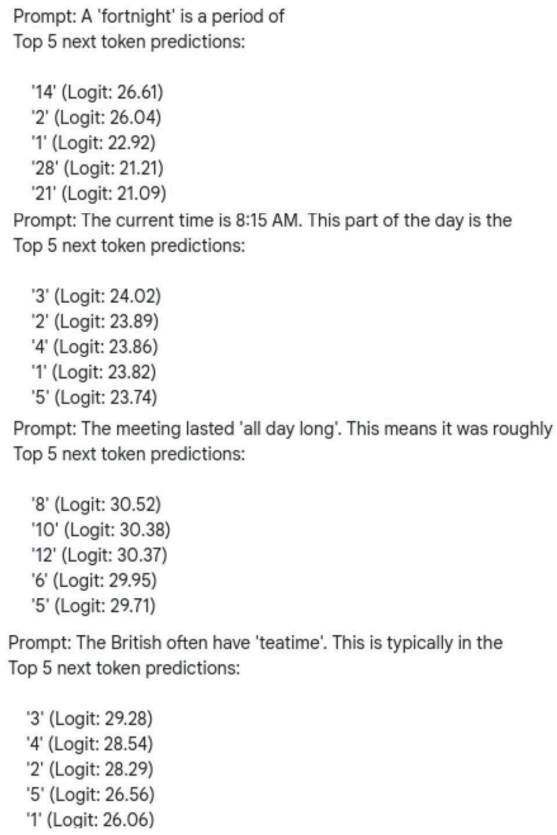
I explored how llama 3.1 1b "**understands**" time. It recalls time facts (fortnight=14d) & does simple arithmetic (8am+3h=11 am). grasps teatime (as pm) & jiffy (as quick). but struggles to link 8:15 am to morning (predicting numbers instead).

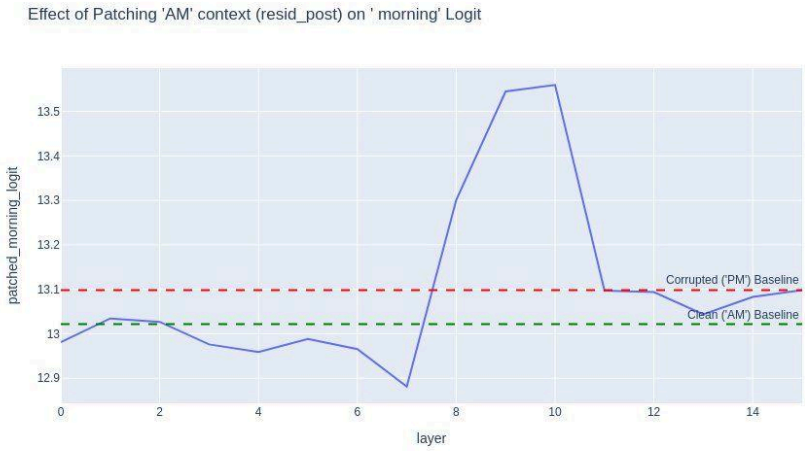
Activation from a clean "9:00 am" context was patched into a corrupted "9:00 pm" context at each layer and effect on the logit of the "morning" token was measured ("the time displayed is 9:00 am. therefore, it is currently the....").

Activation patching showed that outputs from layers 8-10 in an "am" context can make the model predict "morning" even in a "pm" setting, suggesting that these layers are imp for handling am vs pm meaning.

A few neurons that consistently appeared across different examples.

> `N1810@L10`, activated across prompts involving morning & evening.
> `N4988@L14`, `N26@L14`, `N2260@L15`, these neurons tend to activate when the model is producing numeric answers or structural completion.


Prompt: A 'fortnight' is a period of
Top 5 next token predictions:
'14' (Logit: 26.61)
'2' (Logit: 26.04)
'1' (Logit: 22.92)
'28' (Logit: 21.21)
'21' (Logit: 21.09)
Prompt: The current time is 8:15 AM. This part of the day is the
Top 5 next token predictions:
'3' (Logit: 24.02)
'2' (Logit: 23.89)
'4' (Logit: 23.86)
'1' (Logit: 23.82)
'5' (Logit: 23.74)
Prompt: The meeting lasted 'all day long'. This means it was roughly
Top 5 next token predictions:
'8' (Logit: 30.52)
'10' (Logit: 30.38)
'12' (Logit: 30.37)
'6' (Logit: 29.95)
'5' (Logit: 29.71)
Prompt: The British often have 'teatime'. This is typically in the
Top 5 next token predictions:
'3' (Logit: 29.28)
'4' (Logit: 28.54)
'2' (Logit: 28.29)
'5' (Logit: 26.56)
'1' (Logit: 26.06)


Effect of Patching 'AM' context (resid_post) on ' morning' Logit
patched_morning_logit
13.5
13.4
13.3
13.2
13.1
13
12.9
0
2
4
6
8
10
12
14
layer
Corrupted ('PM') Baseline
Clean ('AM') Baseline